Bayesian Inference for Ordinal Data Using Multivariate Probit Models

Earl Lawrence [*], Derek Bingham [†], Chuanhai Liu [‡], Vijayan N. Nair [§]

Abstract

Multivariate ordinal data arise in many areas of applications. This paper proposes new efficient methodology for Bayesian inference for multivariate probit models using Markov chain Monte Carlo techniques. The key idea for our approach is the novel use of parameter expansion to sample correlation matrices. We also propose methodology for model selection. Our approach is demonstrated through several real and simulated examples.

1 Introduction

Much of the data collected in surveys, opinion polls, and behavioral and social science research are in the form of ordinal or ordered categorial data. These are typically the attitudes of the respondents measured on a Likert scale (Likert 1932) such as: strongly disagree, disagree, neutral, agree, and strongly agree. Researchers in these application areas usually analyze the ordered categorical data by giving scores to the categories (for example, 1-5) and analyzing the scores as continuous data. McCullagh (1980) introduced an alternative analysis using a latent variable model. This treats the ordinal responses as grouped data from an underlying latent variable with the cut points for the groupings viewed as unknown parameters. The parameters of interest as well as the cut points (usually treated as nuisance parameters) can be estimated using maximum likelihood, Bayesian methods, or other techniques (e.g., Agresti 1984; Johnson and Albert 1999; and McCullagh and Nelder 1989). The use of Bayesian methods have become popular because the (unobservable) latent variables can be treated as missing and data augmentation techniques can be used effectively. In this context, probit models, where the latent variables are normally distributed, are especially convenient for computational purposes.

[*]Statistical Sciences Group, Los Alamos National Laboratory, Los Alamos NM 87545
[†]Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby BC Canada
[‡]Department of Statistics, Purdue University, West Lafayette IN 47907
[§]Department of Statistics, University of Michigan, Ann Arbor MI 48109

This paper deals with Bayesian inference for multivariate probit models. Ordinal data collected in most surveys are multivariate in nature: multiple item responses or longitudinal responses to the same question or a combination of both. We describe several applications in the next section. In this article, we develop Bayesian inference using Markov chain Monte Carlo techniques to analyze multivariate probit data. An important technical challenge arises from the fact that the covariance matrix of the latent multivariate data is not identifiable, so one has to restrict attention to correlation matrices. Bayesian inference for arbitrary correlation matrices is known to be a difficult problem. Our approach uses a novel idea by relying on the parameter expansion technique (see Liu, Rubin, and Wu 1998, and Liu and Wu 1999). This procedure was originally proposed to accelerate EM and data augmentation algorithms by adding an unidentified parameter to the model. In our setting, we use parameter expansion not for speeding up our estimation procedure (though it does converge quite quickly), but instead to overcome the inherent difficulty in the multivariate probit that arises from parameter identifiability. Using parameter expansion for this case was mentioned in Liu (2001).

There has been other work in the literature on multivariate probit models. Ashford and Sowden (1970) restricted attention to the bivariate case due to computational limitations, and they suggested carrying out all pairwise analyses for more than two categories. Ochi and Prentice (1984) and others have extended the model to higher dimensions for the equicorrelated case. This simplifies the problem by dealing with just a single correlation value but the approach does not extend to a general correlation matrix. Bock and Gibbons (1996) used a parametrization based on matrix decomposition that reduces the number of elements that need to be estimated in the correlation matrix. Again, the restriction makes the problem more computationally tractable but does not allow for a general correlation matrix in the formulation. Chib and Greenberg (1998) described a method for estimation of a multivariate probit model using an MCMC algorithm and a Monte Carlo EM algorithm. This approach is quite general but complicated since each correlation coefficient is drawn separately requiring its own prior specification and Metropolis step. Lastly, a similar approach to the one presented in our paper was recently proposed by Edwards and Allenby (2003). They also implement a MCMC procedure in which the standard identifiability restrictions are ignored, although they do not cast it in the context of parameter expansion. In their case, the completed chain is post-processed via a transformation of each set of parameters. As we shall see, post-processing the completed chain leaves open the chance of numerical instability and a lack of convergence.

In this article, we propose new Bayesian methodology for inference for multivariate probit models. Our approach has a few key features. Firstly, we use parameter expansion as a way to draw correlation matrixes. Secondly, for correct choice of prior distributions, we can draw the correlation parameters independently of the other parameters, thereby reducing the dependence between iterations. Finally, Bayesian variable selection using MCMC is adapted for use in the current setting. The paper is organized as follows. Three applications of multivariate ordinal data are described in Section 2 and the data are analyzed in Section 6. The model and some of the difficulties with parameter estimation are then described. Section 4 describes our methodology for inference with multivariate probit models. The paper concludes with a brief discussion of future work.

2 Motivating Examples

2.1 Application 1: Study on Breast Cancer Prevention

This application is part of on-going research being conducted by the Program for Improving Health Care Decisions, a research group within the Division of General Internal Medicine at the University of Michigan. It is part of a multi-phase effort aimed at understanding how women can make informed decisions on whether or not to take tamoxifen prophylaxis for prevention of breast cancer. One of us has been involved in the statistical design and analysis for this project. This example is representative of multivariate ordinal data collected in many other areas, including marketing, opinion polls, and behavioral and social science research.

Tamoxifen is a synthetic hormone pill used to treat breast cancer as well as reduce the chance of getting breast cancer for women at high risk. This study is concerned with women in the latter category. To make an informed decision, women need to understand their baseline risk of breast cancer as well as the risks and benefits of tamoxifen prophylaxis. The main goal of the overall project is to identify the best ways to communicate such information.

Some of the decision aids (communication methods) that are being studied include the level of presentation detail and the use of different types of graphics for communicating risks. In addition, the effect of the order of presentation of information in the survey instrument (risk before/after benefit; risk perception before/after knowledge) is also of interest. Finally, framing questions in terms of gain versus loss, is also important. The effects of these were studied in a pilot experiment.

The original survey instrument had 38 questions covering a broad range of topics, with most of

the responses being ordinal in nature. A subset of the survey instrument with selected questions, including questions on demographic information, is given in Appendix A. The responses to these selected questions (items) are multivariate ordinal data and are analyzed in Section 6. Questions of interest include the effects of age, education, and income on the various responses.

One could of course analyze the responses to each item separately and ignore the correlation structure. However, treating them as multivariate data leads to a more efficient analysis. More importantly, researchers are interested in the correlations among the various items as these are indications of commonality among the questions. In fact, it is common in the psychometric literature to determine intrinsic dimensions among the large number of items using factor analysis. Thus, estimating the correlation matrix of the latent multivariate data is useful in such studies.

2.2 Application 2: Yield Data in Integrated Circuit Fabrication

The fabrication of integrated circuits (IC's) is a complex process, involving many separate steps and lasting up to 3 months. Several hundred IC's, or chips, are fabricated simultaneously on a *wafer*, the exact number depending on the particular technology. The wafers themselves are processed together in groups called *lots*. Upon completion of the fabrication process, each chip on every wafer is subjected to a number of functionality tests using a probe tester. The probes make voltage and current measurements or verify the operation of the device by observing a predefined response. Typically, tests are categorized in several broad categories: (i) parametric tests which measure electrical properties of pin electronics; and (ii) functional tests which use input vectors and corresponding responses designed to check proper operation of a verified design. The test responses can be either in the form of pass/fail for some categories and ordinal data corresponding to below lower specification level, within specifications, or above specifications. Due to the volume of the data, the exact measurements are not recorded and only the ordinal data are available for analysis.

The traditional analysis of probe data from wafer maps has been restricted to simple summaries such as proportion of defects in the various categories. For process control purposes, one commonly tracks the yield (proportion of "good" chips at the wafer or lot level, with a "good" chip being one that has passed all the tests). However, the defects in the chips are often spatially clustered, with different spatial patterns pointing to different types of process problems. Hansen, Nair, and Friedman (1997) and Friedman, Hansen, Nair, and James (1997) developed methods for exploiting the spatial information in wafer map data for yield and process improvement. Their methods used only binary data (good chips versus failed chips). Kutsyy (2001) considered latent variable models

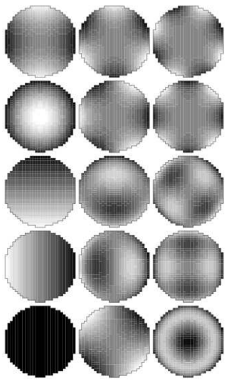

Figure 1. Spatial wafer bases for fault detection.

for ordinal data using a Markov random field model for the latent process.

Here, we consider multivariate ordinal data corresponding to different categories for each of the tests and treat the spatial aspect of the problem through a regression model. Specifically, Figure 1 depicts some common types of spatial patterns that arise due to process problems during fabrication. These are actually a subset of Zernike polynomials which form an orthonormal basis function on the unit disk. In Section 6, we describe how one can fit regression models for the selected basis functions with the active regression coefficients indicating the nature of the process problem in the data.

2.3 Application 3: Six Cities Example – Longitudinal Study on Air Pollution

The third example is a longitudinal study on the health effects of air pollution (see Ware, Dockery, Spiro III, Speizer, and Ferris Jr. 1984). Chib and Greenberg (1998) considered a subset of the data, containing repeated measurements of wheezing status of 537 children in southern Ohio with a yes/no response regarding the occurrence of wheeze at ages 7, 8, 9, and 10. The covariates of interest were the child's age, centered at 9, and an indicator variable for the mother's smoking habit during the first year of the study. Although only binary data are involved in this example, we will use it to demonstrate the usefulness of the methods developed here for analyzing multivariate binary data and to compare our results with other analyses in the literature.

In this example, the correlation matrix in the multivariate set up provide a measure of how the relationship is changing with time. By estimating the correlation matrix and examining the structure, we can determine if the data follow some special structure such as an AR or ARMA models in the latent variables. This would allows us to characterize the nature of the temporal relationship in the longitudinal responses.

3 The Models

3.1 Univariate Probit Model

We begin with a brief introduction to the univariate probit (UVP) model. For more details, see Agresti (1984) or Johnson and Albert (1999).

Let $\{Y_i,\ X_i\}$ $i = 1, \ldots, n$ be the i^{th} observation, where each Y_i can take one of k ordered categorical values coded $1, \ldots, k$ and X_i is a vector of covariates. For each Y_i, let Z_i be a latent normal variable with mean $X_i\beta$ and variance, σ^2, set to unity (see below). Let γ be a vector of cutpoints that define the relationship between response, Y_i, and the latent variable, Z_i. Specifically, if $\gamma_{c-1} < Z_i \leq \gamma_c$ $(c = 1, \ldots, k)$, then $Y_i = c$, with $\gamma_0 = -\infty$ and $\gamma_k = \infty$.

This model results in the probability distribution for Y

$$P\{Y_i = c\} = \Phi(\gamma_c - X_i\beta) - \Phi(\gamma_{c-1} - X_i\beta). \tag{1}$$

Inference for this UVP model involves estimation of the regression coefficients β and the cutpoints γ.

There are some identifiability issues associated with this model. First, the observed ordinal data are invariant up to monotone transformations of the underlying latent distribution. In particular, this implies that the variance of the Z_i's cannot be estimated. To see this, suppose instead that an unrestricted variance, σ^2, is allowed. We can then form new regression coefficients $\alpha = \sigma\beta$ and new cutpoints $\theta = \sigma\gamma$ that result in the same observed probabilities given in (3.1). Consequently, an unrestricted variance results in an unidentifiable model. Thus, for identifiability reasons, it is common to restrict the variance $\sigma^2 = 1$. Second, one cannot estimate all of the cut-points in addition to the overall intercept term in the regression. To resolve this issue, it is common to set the first cut-point $\gamma_1 = 0$.

Maximum likelihood estimation of this model was first discussed in McCullagh (1980). See also Agresti (1984) where a Newton-Raphson routine is used for parameter estimation. Bayesian estimation is discussed in Johnson and Albert (1999), where both the choice of priors and the implementation of an MCMC algorithm are introduced. Unfortunately, these univariate methods of inference do not extend in a straightforward manner to the multivariate setting.

3.2 The Multivariate Probit Model

Consider now the modeling of multivariate ordinal data. Denote the i^{th} observation by $\{Y_i, X_i\},\ i = 1, \ldots, n$, where each Y_i is q-vector in which each entry $Y_{i,j}$ $(j = 1, \ldots, q)$ can take one of k_j ordered

values coded $1, \ldots, k_j$. Further X_i is a $p \times 1$ vector of predictors. As before, let Z_i be the latent normal variable for each observed Y_i. Now the latent variable is q-variate normal with the vector of means determined by the predictors and regression parameters. Specifically, let β_j be a $1 \times p$ vector of regression coefficients associated with the j-th response. The mean for $Z_{i,j}$ is given by $\beta_j X_i$. If the matrix of regression coefficients is denoted by $\beta = (\beta_1', \beta_2', \ldots, \beta_q')'$, then Z_i is distributed normally with mean βX_i.

Once again, the relationship between Y_i and Z_i is defined by a set of cutpoints γ, one set for each ordinal response. In particular, if $\gamma_{j,c-1} < Z_{i,j} \leq \gamma_{j,c}$, then $Y_{i,j} = c$, with $\gamma_{j,0} = -\infty$ and $\gamma_{j,k_j} = \infty$ for all j. As before, we set $\gamma_{j,1} = 0$ for all j to resolve identifiability problem with the intercepts.

As in the univariate case, there is also an identifiability problem associated with the variance-covariance matrix of the Z_i's. Since the ordinal data are invariant under monotone transformations of the underlying latent variables, the variance-covariance matrix can be estimated only up to scaling constants. Thus, for identifiability purposes, we have to restrict the variance-covariance matrix to be a correlation matrix R. This is analogous to setting $\sigma^2 = 1$ in the univariate case, but this essentially removes the parameter σ^2 in the univariate framework. In the multivariate case, we are restricting the structure of the covariance matrix (i.e., all diagonal elements equal one). This presents the main challenge for multivariate probit models as it is difficult to do Bayesian inference for general correlation matrices.

4 Multivariate Probit Inference

We make use of the matrix-variate normal distribution in order to simplify the algebra and presentation (though it also eases implementation). Now, let M and μ be $q \times n$ matrices, and let $\text{vec}(M)$ be the vector resulting from stacking the columns of M. Also, let Σ be a $q \times q$ matrix and Ψ be an $n \times n$ matrix. The matrix M is distributed as a matrix-variate normal variable with mean matrix μ and covariance matrix $\Sigma \otimes \Psi$ if $\text{vec}(M')$ is distributed as a qn-variate normal variable with mean vector $\text{vec}(\mu')$ and covariance matrix $\Sigma \otimes \Psi$. We denote this distribution $M \sim N_{q,n}\{\mu, \Sigma \otimes \Psi\}$. See Gupta and Nagar (2000) for a more complete description of the matrix variable normal density.

To take advantage of this representation for our model, we start by constructing the following matrices from the latent and covariate vectors, respectively: $Z = [Z_1, Z_2, \ldots, Z_n]$ and $X = [X_1, X_2, \ldots, X_n]$ where Z_i is the q-vector of latent variables and X_i is the p-vector of co-

variates for response unit i. Now Z is matrix-variate normal with mean matrix βX and covariance matrix $R \otimes I_n$. The resulting complete data likelihood for the multivariate probit model is

$$\mathcal{L}(\beta, R, \gamma) = |R|^{-\frac{n}{2}} \text{etr}\{-\frac{1}{2} R^{-1}(Z - \beta X)(Z - \beta X)'\} \prod_{i=1}^{n} \prod_{j=1}^{q} \mathcal{I}\{\gamma_{j,Y_{i,j}-1} < Z_{i,j} \leq \gamma_{j,Y_{i,j}}\}. \quad (2)$$

This is simply the likelihood based on the matrix-variate normal density with each component of the random matrix restricted to particular region based on the observed values in Y.

The key feature of the likelihood in (2) is the truncation that arises from the fact that given Y_i, we do not know Z_i exactly; we only know the rectangular region that contains it. The observed data likelihood is obtained by integrating out the latent variable Z over the ranges specified by the truncation. The resulting likelihood is difficult to maximize directly. The presence of the missing latent variables suggests using the EM algorithm to impute the missing data and maximize the more easily manipulated complete data likelihood. Again, the truncation makes the imputation difficult as there is no closed form for the for the expectation of the latent variables. Other numerical optimization techniques are similarly hindered by the difficulty of computing integrals and derivatives of this quantity.

Simulation-based techniques are known to be computationally efficient and easily tractable for this type of setting. Straightforward Monte Carlo integration is not feasible due to the difficulty of producing samples from a truncated multivariate normal distribution. It is possible to simulate draws from such a distribution using MCMC. The conditional distribution of each component of a truncated multivariate normal variable is simply truncated univariate normal. By cycling through the conditional draws, we can obtain a Markov chain whose stationary distribution is the required truncated multivariate normal.

Because of the need to use MCMC to generate the latent variable, we will frame the whole problem in the Bayesian MCMC setting. In the next section, we consider the use of conjugate priors for the model parameters that allows for a fairly straightforward MCMC implementation (i.e., normal priors for the regression coefficients, uniform priors for the cutpoints and an inverse Wishart prior for the covariance matrix). An attractive feature of this approach is that each parameter draw is a Gibbs step from a relatively simple distribution. Further, additional techniques explained in the next section ensure that the chain mixes quickly. The result is an algorithm that is simple to implement and efficient to run.

A challenging obstacle arises, however, when the covariance matrix is restricted to a correlation matrix. It is convenient to use the inverse Wishart distribution to draw unstructured covariance

matrices. However, this cannot easily be modified to produce draws that have ones on the diagonal. The algorithm would be greatly simplified if we can take general inverse Wishart draws and then transform the parameters so that the restriction is met. Surprisingly, we can do exactly this.

4.1 Expanding the Parameters

The methodology uses an expanded parameter formulation (Liu et al. 1998 and Liu and Wu 1999) which enables estimation of the restricted covariance matrix from a single draw from an inverse Wishart distribution.

Consider our observed data model:

$$\mathrm{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} =$$
$$\int_{\gamma_{q,y_q-1}}^{\gamma_{q,y_q}} \ldots \int_{\gamma_{1,y_1-1}}^{\gamma_{1,y_1}} (2\pi)^{-\frac{q}{2}} |R|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(Z-\beta X)' R^{-1} (Z-\beta X)\} dZ_1 \ldots dZ_q.$$

Let the matrix V be a $q \times q$ diagonal matrix with positive diagonal elements $v_1, \ldots, v_q$. Now consider the following scale transformation on the latent variable: $W = V^{\frac{1}{2}} Z$. Make this substitution into the above observed data model. After some tedious algebra, we arrive at

$$\mathrm{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} =$$
$$\int_{\sqrt{v_q}\gamma_{q,y_q-1}}^{\sqrt{v_q}\gamma_{q,y_q}} \ldots \int_{\sqrt{v_1}\gamma_{1,y_1-1}}^{\sqrt{v_1}\gamma_{1,y_1}} (2\pi)^{-\frac{q}{2}} |V|^{-\frac{1}{2}} |R|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(W - V^{\frac{1}{2}}\beta X)' (V^{\frac{1}{2}} R V^{\frac{1}{2}})^{-1} (W - V^{\frac{1}{2}}\beta X)\}$$
$$dW_1 \ldots dW_q.$$

This is a useful formation because the probabilities for Y remain the same despite the change of variable. If we rewrite the parameters in the following way:

$$\begin{aligned} \alpha &= V^{\frac{1}{2}}\beta, \\ \theta_{j,c} &= \sqrt{v_j}\gamma_{j,c}, \\ \Sigma &= V^{\frac{1}{2}} R V^{\frac{1}{2}}, \end{aligned}$$

the above becomes the probability for Y based on the general multivariate normal density with no restriction on the variances. That is,

$$\mathrm{P}\{Y_1 = y_1, \ldots, Y_q = y_q\} =$$
$$\int_{\theta_{q,y_q-1}}^{\theta_{q,y_q}} \ldots \int_{\theta_{1,y_1-1}}^{\theta_{1,y_1}} (2\pi)^{-\frac{q}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(W-\alpha X)' \Sigma^{-1} (W-\alpha X)\} dW_1 \ldots dW_q. \qquad (3)$$

We call this the expanded parameter model because the set of parameters has grown to include a set of variances which were previously fixed at one. This leads to the following expanded likelihood:

$$\mathcal{L}(\alpha, \Sigma, \theta) = |\Sigma|^{-\frac{n}{2}} \text{etr}\{-\frac{1}{2}\Sigma^{-1}(W - \alpha X)(W - \alpha X)'\} \prod_{i=1}^{n} \prod_{j=1}^{q} \mathcal{I}\{\theta_{j,Y_{i,j}-1} < W_{i,j} \leq \theta_{j,Y_{i,j}}\}. \quad (4)$$

This model matches the observed data model, but does not meet the identifiability restrictions. In other words, the observed data do not allow us to estimate the parameters as they are described in 3. However, if we could use this model, we would be able to define an MCMC scheme that used a very simple inverse Wishart posterior for the covariance matrix. We propose making draws from this model and then transforming them to meet the identifiability conditions.

This scheme has a number of nice features. First, all of the draws are relatively simple. There is no complicated attempt to individually draw each correlation parameter; just one draw of an unrestricted covariance matrix. Secondly, for correct choice of prior distributions, we can draw Σ independently of the other parameters and reduce dependence between iterations. In other words, for each iteration of the overall MCMC scheme, the entire parameter draw is based only on the last latent variable draw. Both of these features will improve the mixing rate of the MCMC algorithm and result in improved performance. Additionally, the theory of parameter expansion guarantees a further gain in convergence rate based on the implicit resampling of the variances and the adjustment to the parameters. Next we will focus on the implementation of this model and then we will discuss the related convergence issues.

4.2 Priors

We focus here on noninformative conjugate priors for two reasons. First, in most cases, we prefer to let the data decide the parameters with little or no influence from outside sources. Secondly, the noninformative priors lead to convenient results regarding the draws and convergence properties. Our experience has shown that the procedure will also work with other priors but general results are not easy to obtain.

The prior chosen for the covariance matrix is the Jeffreys' noninformative prior for scale matrices. This prior,

$$\pi(\Sigma) \propto |\Sigma|^{-\frac{q+1}{2}}, \quad (5)$$

leads to an inverse Wishart full conditional distribution for the covariance draw. A key feature is that it keeps the expanded scale parameters independent from the correlations *a priori* since

$$|\Sigma|^{-\frac{q+1}{2}} = |V|^{-\frac{q+1}{2}} |R|^{-\frac{q+1}{2}}.$$

We choose a matrix-variate normal prior for the regression parameter matrix β. Recall that each row of the β matrix contains the regression parameters for one dimension of the response. With this in mind, we use the following prior formulation:

$$\pi(\beta) = N_{q,p}(B, \mathcal{B} \otimes \mathcal{W}). \tag{6}$$

The matrix B represents the prior belief about the location of β. The matrices $\mathcal{B}$ and $\mathcal{W}$ represent the prior belief about the relationship between the regression coefficients. The matrix $\mathcal{W}$ contains prior information regarding the relationship among regression coefficients within each dimension and the matrix $\mathcal{B}$ contains the prior information regarding the relationship among regression coefficients across dimensions. There are a few things to note about the choice of prior here. First, the development in section 4.3 is done by assuming that $\pi(\beta) \sim 1$ for a noninformative prior. Second, if we let the prior for β depend on R by setting $\mathcal{B} = R$, we can integrate out β allowing us to draw the covariance independently of β for an efficient MCMC implementation. This property also applies to the noninformative prior. Lastly, since the matrix variate normal distribution can be rewritten as a multivariate normal distribution, the prior for beta can be specified in a more general fashion. This issue will be further developed in the upcoming section on variable selection.

Finally, we choose normal priors with an order restriction for γ. In other words, for a particular dimension, let all of the free cutpoints follow an independent multivariate normal distribution with the restriction that $\gamma_{j,1} = 0$, $\gamma_{j,2} \geq \gamma_{j,1}$, etc. Note also that the order restriction has no practical outcome as the likelihood will impose this restriction naturally. A noninformative prior is achieved by allowing the variances to go to infinity.

4.3 MCMC Implementation

We now consider the posterior distributions and the steps of the of MCMC scheme. We start with a brief description of the latent variable draw, followed by consideration of the other parameters.

Each latent observation consists of a q-variate normal variate that is truncated along each dimension. Consider a variable Z_i with mean $\mu_i = \beta X_i$ and covariance matrix R. Let this variable be truncated in all dimensions with a set of cutpoints so that $\gamma_{j,a} < Z_{i,j} \leq \gamma_{j,b}$ $(a < b)$. The univariate conditional distributions are all univariate truncated normal where the parameters are simple functions of μ_i and R. Specifically,

$$\begin{aligned} \sigma^2 &= 1/R^{-1}_{j,j}, \text{ and} \\ \nu &= \mu_{i,j} - \sigma^2 (R^{-1})_{j,-j} (Z_{i,-j} - \mu_{i,-j}). \end{aligned}$$

Thus, we can create a Gibbs' sampler to cycle through the conditionals to produce a Markov chain with the desired truncated multivariate normal distribution. In practice, for this algorithm, we choose some small fixed number of iterations to run the chain in order to produce each draw.

We now treat the data as having arisen from the expanded model. Given the latent variables W, we turn next to the covariance matrix. As previously noted, we are going to draw this as if it were unrestricted. Because of the noninformative choice for the beta prior, this draw is made independently of the other variables, conditional only on the missing data, which allow us to make the entire parameter draw conditional on only the latent variables. Combining our prior (5), expanding the likelihood (4), and integrating out α gives

$$\Sigma|Z \sim IW\{n-p+q-1, WW' - WX'(XX')^{-1}XW'\}.$$

Next, we move on to the regression coefficients. Using the noninformative prior for β (6, with mean 0 and infinite variances) and combining it with our expanded likelihood (4), we see that we can draw from the expanded model:

$$\alpha|W, \Sigma \sim N_{q,p}\{WX'(XX')^{-1}, \Sigma \otimes (XX')^{-1}\}.$$

Finally, we draw the cutpoints. Using the noninformative prior achieved by letting the variances go to infinity, we see that:

$$\theta_{j,c}|W, Y \sim Uni(\max_i\{W_{i,j}|Y_i, j=c\}, \min_i\{W_{i,j}|Y_{i,j}=c+1\})$$

for all freely varying $\theta_{j,c}$. If we use an informative prior based on the normal distribution, then each posterior draw is normal truncated to the region defined for the Uniform draw.

With the current draw from the expanded model, we simply remove the expanded parameters via re-scaling to get the desired sample of interest. We use the following transformation:

$$\begin{aligned} Z_{i,j} &= W_{i,j}/\sqrt{\Sigma_{j,j}}, \\ R_{i,j} &= \Sigma_{i,j}/\sqrt{\Sigma_{i,i} \times \Sigma_{j,j}}, \\ \beta_{k,j} &= \alpha_{k,j}/\sqrt{\Sigma_{j,j}}, \text{ and} \\ \gamma_{j,c} &= \theta_{j,c}/\sqrt{\Sigma_{j,j}}. \end{aligned} \tag{7}$$

In summation, we have the following algorithm for the MVP-MCMC algorithm starting with some set of latent variables and parameters $(Z^{(r-1)}, R^{(r-1)}, \beta^{(r-1)}, \gamma^{(r-1)})$.

1. Draw $W^{(r)}|Z^{(r-1)}, R^{(r-1)}, \beta^{(r-1)}, \gamma^{(r-1)}$.

2. Draw $\Sigma^{(r)}|W^{(r)}$.

3. Draw $\alpha^{(r)}|W^{(r)}, \Sigma^{(r)}$.

4. Draw $\theta^{(r)}|Y, W(r)$.

5. Transform according as in (7)

Remark: This algorithm requires us to start with an initial guess for the latent variables as well as guesses for the parameters. This additional requirement is due to the nature of the procedure that produces the latent variables. Some experience suggests that a poor choice of starting values can be difficult to overcome. In order to simplify this process, we recommend running a few iterations of the UVP model on the marginal data. The UVP model does not require that the latent data be initialized. Running several iterations of the UVP algorithm for each dimension results in a set of starting points that fit together, (i.e., the latent variables are between the cutpoints and are drawn from the regression coefficients). This process amounts to choosing a starting point based on the assumption of independence among the dimensions and seems to work well in practice.

4.4 Convergence Under Parameter Expansion

Consider the standard data augmentation setting. Let the observed data be y and the complete data be z. Let the model parameters be represented by θ. The basic components are the observed-data model, $f(y|\theta)$, and the complete-data model, $f(y, z|\theta)$. These models have the following relationship:

$$f(y|\theta) = \int f(y, z|\theta)dz. \tag{8}$$

Additionally, we have a prior for the parameters, $\pi(\theta)$. The approach proceeds by first drawing $z \sim f(z|y, \theta) \propto f(y, z|\theta)$ and then drawing $\theta \sim f(\theta|y, z) \propto f(y, z|\theta)\pi(\theta)$. This process repeats until the Markov chain converges to the stationary distribution whereupon we can use the sample of θ to obtain estimates of the quantities of interest.

Suppose that there is a parameter that can be identified in the complete-data model, but which is unidentifiable from the observed data alone. Call this parameter α. We can now expand the complete-data model to $f(y, w|\theta, \alpha)$ and still preserve the observed data model. In mathematical terms, we have

$$\int p(y, w|\theta, \alpha)dw = f(y|\theta). \tag{9}$$

Let the expanded parameter set have prior $\pi(\theta, \alpha)$. If the marginal prior for θ resulting from this joint distribution is the same as the prior from the original data augmentation scheme, then the posterior for θ from the expanded model will be the same as the posterior for θ from the original model. In other words, if we preserve the prior for our regular parameters, then expanding the parameter set will not affect the posterior for our regular parameters. A more detailed explanation can be found in Liu and Wu (1999) where they provide several examples of parameter-expanded data augmentation (PX-DA) schemes.

By choosing priors as we have done, we achieve the above property. The posterior distributions for our parameters are unchanged by expanding the parameters. The model and scheme presented here coincides with Scheme 2 of Liu and Wu (1999). We draw the latent variable from a set of parameters that meet the usual restriction. We follow this with a draw of both the regular and expanded parameters from the expanded model. The regular parameters are then used for the next latent variable draw. Liu and Wu (1999) present the theoretical background for convergence in some detail. We simply offer some intuition into the convergence issues using their idea of orbits.

For any MVP model, there are an infinite number of parameter sets that give the same observed model: the latent variable can be scaled along any and all dimensions and the appropriate scaling of the parameters will not change the model. Call this infinite set of models the orbit. Any model in an orbit is equivalent to any other model in that orbit. In other words, the orbit determines the observed model and the location within the orbit does not matter. Each iteration of the MCMC algorithm chooses an orbit and a location within that orbit. The transformation simply moves the model parameters to the place in the orbit that meets our requirements. It does not choose a different orbit. Thus, the MCMC scheme can be imagined in the usual manner alternating between data augmentation and model selection.

Remark: The authors of the parameter expansion papers developed the method in order to gain speed. In the EM setting, parameter expansion uses covariance adjustment to speed the convergence rate of the algorithm. In the data augmentation setting, the covariance adjustment hastens the convergence to the stationary distribution. Both these goals are achieved by introducing an unidentified parameter. In our setting, we instead use the idea of parameter expansion to simplify the estimation procedure and overcome model non-identifiability.

4.5 Prediction

An important part of using a statistical model is prediction. For example, consider a bivariate example in which a product undergoes a series of pass/fail tests. The probability of any failures gives the overall failure rate of the product. Alternatively, consider a trivariate example in which a product undergoes a series of test to determine if it is below, within, or above specs. The probability of being within specifications on all the test is obviously of interest. These probabilities are available for each iteration of the Markov chain.

At each iteration, we have a complete set of parameters: regression coefficients, covariance matrix, and cutpoints. With this information, there are several ways that we can determine various probabilities. The simple approach is to estimate the probabilities via simulation. With a given set of parameters and covariates, multivariate normal variables can be simulated and the probabilities estimated. Embedding this into the MVP MCMC has some attractive properties. This will give us a sample of desired probabilities that can be used to form point estimates or sample-based confidence intervals. For more sophisticated techniques on evaluating the probabilities, see Gassmann, Deák, and Szánti (2002).

5 Variable Selection

A complete analysis for this type of model would naturally include variable selection. The stochastic search variable selection of George and McCulloch (1993) lends itself well to inclusion in the procedure discussed here. The key ingredient of their algorithm is the inclusion of an indicator variable λ_i describing the presence or absence of a covariate i in a regression model. The prior for the regression coefficients depends on the indicator variable. If $\lambda_i = 1$, then β_i is normally distributed with large variance and mean zero, allowing large values. Otherwise, the prior for β_i is normally distributed around zero with small variance. We now develop this idea for the current setting. First, we introduce a matrix of indicator variables λ where $\lambda_{j,k} = 1$ if the model for response j includes the covariate k. This parameter will be included in our MCMC procedure. The prior for β will be conditional on λ. For the present development, we will treat the coefficients as independent *a priori*. Thus, $\text{Var}(\beta_{j,k}|\lambda_{j,k}) = c_{j,k}^{2\lambda_{j,k}}\tau_{j,k}^2$, where $\tau_{j,k}$ is small and $c_{j,k}$ is large to produce the desired effect. For this development, we use the automatic tuning parameter values described in Chipman (1998). We set $c_{j,k} = 10$ to indicate an order of magnitude difference between significant and insignificant effects. We set $\tau_{j,k} = 1/[3 \times \text{range}(X_k)]$ for the reasons discussed in

Chipman (1998). The combined prior for β is such that $\text{vec}(\beta')$ is qp-multivariate normal with mean $\vec{0}$ and $\mathcal{D}$, a diagonal covariance matrix with the entries above. For λ, we use the independent prior with $\text{P}\{\lambda_{j,k} = 1\} = p_{j,k}$, *a priori*. All other priors remain the same, but the posteriors will change.

Because of the form of the prior for β, it cannot be integrated out of the likelihood and the draw of the covariance will be dependent on the regression parameters. Thus, combining the likelihood with the priors, we have the following posteriors:

$$\begin{aligned}
\Sigma|Z,\beta &\sim IW\{n+q+1,(Z-\beta X)(Z-\beta X)'\} \\
\text{vec}(\alpha')|Z,\Sigma &\sim N_{qp}\{\mathcal{V}\mathcal{M},\mathcal{V}\} \\
&\quad \mathcal{V}=[\mathcal{D}^{-1}+(\Sigma^{-1}\otimes XX')]^{-1},\ \ \mathcal{M}=(\Sigma^{-1}\otimes XX')\text{vec}[(XX')^{-1}XZ'] \\
\text{P}\{\lambda_{j,k}=1|\beta\} &= \frac{a}{a+b}
\end{aligned}$$

where $a = \pi\{\beta|\lambda_{(j,k)},\lambda_{j,k}=1\}p_{j,k}$ and $b = \pi\{\beta|\lambda_{(j,k)},\lambda_{j,k}=0\}(1-p_{j,k})$. The draws for the cutpoints are unchanged. The transformation is also unchanged, but no transformation is required for λ.

Cycling through the draws in the order described will produce a Markov chain that includes the usual parameters and a sample of λ draws that can be used to perform variable selection. Each λ_i represents a model choice. The models occurring most frequently indicate those variables that should be included in the model. For more details about the overall procedure and analysis of the results, please refer to George and McCulloch (1993). The procedure will be considered in the wafer and breast cancer survey examples below.

The above posteriors are typical of what is seen when the prior for β differs from the flat prior and the parameter cannot be integrated out of the posterior. The draws are still relatively simple and from well-known distributions, but the correlation between consecutive draws is increased by the dependence of Σ on the last set of parameters.

6 Analysis of Data from the Examples

6.1 Six Cities – Air Pollution Longitudinal Data

We will consider the subset of the data also analyzed in Chib and Greenberg (1998). These are repeated measurements of wheezing status of 537 children in southern Ohio with a yes/no response regarding the occurrence of wheeze at ages 7, 8, 9, and 10. The covariates are the child's age,

centered at 9, and an indicator variable for the mother's smoking habit during the first year of the study.

Chib and Greenberg (1998) fit a single linear model for the response, with the child's age appearing as both a measurement variable and also a covariate. That is,

$$\mathrm{E}\{z_{i,j} = 1\} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3}$$

where the covariates are age, smoking indicator, and their interaction.

These data are quite attractive in our setting because they allow for fitting of the full multivariate probit model, with a separate regression model for each response variable. That is,

$$\mathrm{E}\{z_{i,j} = 1\} = \beta_{j,0} + \beta_{j,1} x_{i,1}$$

where $j = 1, \ldots, 4$ and the covariate is the smoking indicator. Because of differences in formulation, the specific forms of the models will differ between our presentation and theirs, but the results can still be compared. Our model has four linear components for each response with coefficients for the intercept and smoking habit. Each of our intercepts will be compared with their intercept plus the adjustment for age and its coefficient. Similarly, each of our smoking coefficients will be compared with their main effect smoking coefficient plus the adjustment for age and the interaction coefficient.

The algorithm was run for 500 iterations. Plots of the Markov chains for the regression coefficient matrix and the correlations can be seen in Figure 2. Considering the dimensionality, the chain converged quite quickly even with relatively few observations. Ignoring the first 50 observations as transitionary, we obtain the following parameter estimates:

$$\hat{\beta} = \begin{pmatrix} -.984 & -.00344 \\ -1.027 & .220 \\ -1.052 & .171 \\ -1.241 & .146 \end{pmatrix} \quad \hat{R} = \begin{pmatrix} 1 & .617 & .551 & .593 \\ .617 & 1 & .716 & .593 \\ .551 & .716 & 1 & .666 \\ .593 & .593 & .666 & 1 \end{pmatrix}. \tag{10}$$

Comparing our results to those in Chib and Greenberg (1998), the intercept terms match very well. They estimate the intercept at -1.127, and their age coefficient is estimated at -0.079. The coefficients for smoking do not match up as well, though their estimate for the smoking main effect coefficient, 0.160, is close to the average of our smoking coefficients. Differences arise when we take into account the adjustments made regarding the interaction of age and smoking. Our estimates for the coefficient of smoking appear to change nonlinearly with age. This is comparable to the model fit by Chib and Greenberg where they identify an interaction term, though they only consider an interaction that is linear.

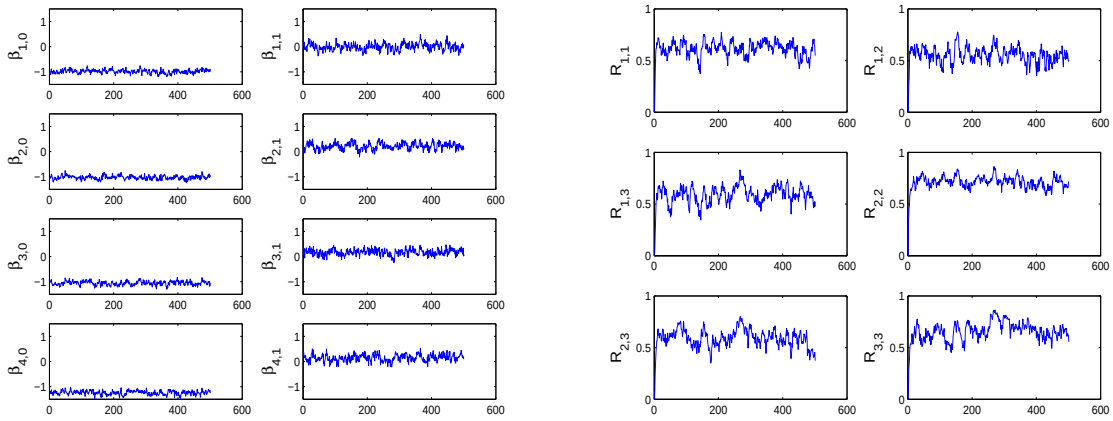

Figure 2. Markov chains of the coefficient matrix and correlations for Chib and Greenbreg's Six Cities Data.

This example highlights some of the important differences between the two procedures. Chib and Greenberg do not provide any plots to visually judge convergence, but they state that 500 burn-in iterations, followed by 10,000 estimation iterations were used to fit the model. In contrast, our chain converges quickly, giving useful results after about 500 iterations. The convergence is relatively fast for several reasons. First, by drawing the correlation matrix as one unit, we avoid the complex conditioning that results from drawing each correlation component partly conditional on the other correlation components. Reducing this dependence speeds the mixing of the chain. Further, in our formulation we draw the correlation matrix conditional on the current values of the augmented data. The rest of the parameters are drawn conditional on the current augmented data and the correlation matrix. The result of this procedure is that the entire parameter draw is conditional only on the current values of the augmented data. Given the current values of the latent variables, the parameter draw is independent of the last parameter draw which again speeds the convergence. Finally, as demonstrated in the previous parameter expansion work (Liu et al. 1998 and Liu and Wu 1999), this procedure naturally reduces dependence between draws and increases the rate of convergence. As an additional note, the procedure is quite easy to implement as each parameter draw is a simple Gibbs step from a well-known distribution.

This example can also be used to contrast our method with the approach described by Edwards and Allenby (2003). They describe an algorithm which is similar in spirit to the one presented here, but with two key differences. One fundamental distinction between the two is when to apply the transformation that forces the parameters to meet the restriction. Referencing work by

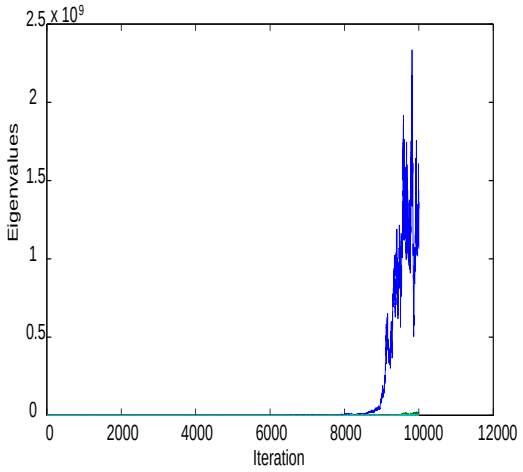

Figure 3. Eigenvalues for each covariance draw from the Edwards and Allenby procedure.

McCulloch and Rossi (1994), their procedure allows the chain to run without restriction, followed by transformation of each parameter set after the chain is complete. Allowing the MCMC to proceed in the unidentifiable model space and correcting the chain by post-processing gives cause for concern in this setting. For example, Figure 3 shows the eigenvalues for each draw of the covariance matrix from the their procedure applied to the current Six Cities example. The values range off dangerously high as the chain continues. We observe that numerical instability can result from the procedure when the data set is small relative to the number of parameters. The issue also occurs in problems requiring long burn-in. The practical problem results from a feedback loop between the imputed latent data and the covariances. As covariances become large, the augmented data is generated with further spread. This augmented data can, in turn, produce larger covariances. When the numerical instabilities arise, it calls into question the convergence of the Markov chain. In our setting, we scale out the expanded parameters at each step of the chain and thus avoid the problem entirely. The second distinction has to do with the nature of the draws. As previously mentioned, our algorithm reduces dependency between draws by integrating out the mean vector before drawing the covariance matrix. The covariance matrix is drawn based only of the data and then the regression coefficients are drawn conditional on the covariance matrix. The result is that the consecutive draws of the entire parameter vector are independent given the augmented data imputed between them. Again, this improves the mixing of the chain.

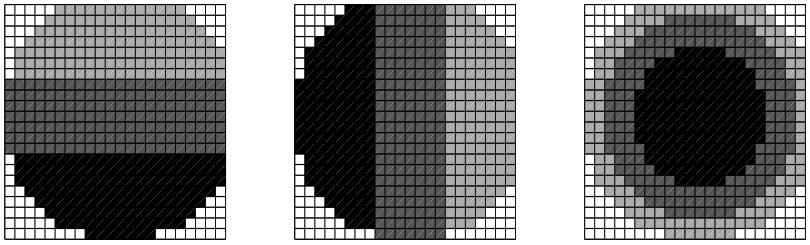

Figure 4. Discrete spatial wafer bases for fault detection in the simulation example.

6.2 Yield Data from IC Fabrication

For this application, we use simulated data to illustrate the methodology. Real data are not available due to propreitary reasons. Moreover, simulated data are useful in illustrating how well the methods do by comparing the estimates to true values of the parameters.

Data were simulated from the wafer map testing procedure described in Section 2. Three covariates corresponding to the three simplified spatial bases in Figure 4 were used. These are discrete versions of three basis functions in Figure 1. The black is set at 1, the dark gray is .5, and the light gray is 0. For example, the chip in row eight, column 1, would have covariates (including intercept term) $X = [1, .5, 1, 0]$. The specific values are unimportant, rather the shape is the key factor. Each basis is supposed to represent a type of process problem. The data were generated from a true model in which the effects of the left-to-right and top-to-bottom gradients were not signficant, but the defects resulting in the inside-out basis was important. We simulate five tests: three binomial tests and two three-level tests. We generated one wafer map with 429 chips. The regression coefficients are given by:

$$\beta = \begin{bmatrix} -1 & 0 & 0 & 1 \\ -2 & 0 & 0 & 1 \\ -1.5 & 0 & 0 & 3 \\ .5 & 0 & 0 & 2 \\ .75 & 0 & 0 & 2 \end{bmatrix}. \tag{11}$$

The last column gives the coefficients for the inside-out basis which had an effect on all five tests. The correlation matrix is

$$R = \begin{bmatrix} 1 & .2 & 0 & 0 & 0 \\ .2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & .5 & .3 \\ 0 & 0 & .5 & 1 & .4 \\ 0 & 0 & .3 & .4 & 1 \end{bmatrix}. \tag{12}$$

Finally, the additional cutpoints for tests 4 and 5 were given by 2.5 and 3.5 respectively.

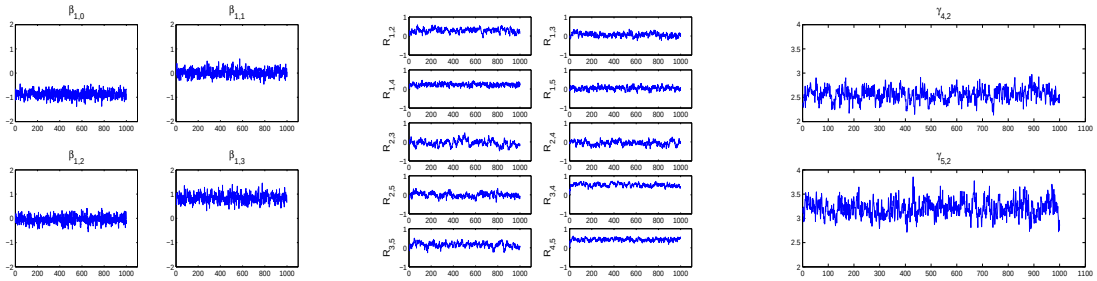

Figure 5. Markov chains of β_1, correlation matrix eigenvalues, and the cutpoints for the wafer example.

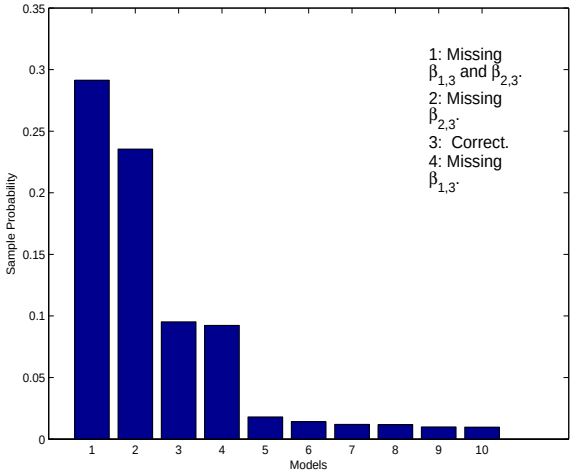

Figure 6. Sample probabilities for the ten most frequent models from the wafer map example. Models 1-4 are the most significant.

In addition to model fitting, determining active spatial basis is a key ingredient in fault detection. In order to pursue this goal, we implement MVP model with SSVS in order to choose significant covariates.

From the data, we produced 1000 iterations from the Markov chain. Figure 5 shows the chains for selected parameters. As before, the chain converges quite quickly from the starting values obtained from the marginal UVP models.

To answer the spatial defect question, consider Figure 6. This figure shows the probabilities for the ten most frequent models in the sample. Models one through four occur much more often than any other. The third most popular model is the correct one from which the data were generated. The most popular model loses $\beta_{1,3}$ and $\beta_{2,3}$ and the other two in the top four leave out one. The true value of each of these parameters is one, which is equal to the error of the latent data causing some uncertainty about their significance. Any examination of these results would likely lead to

further scrutiny of the top four models and their constituent covariates.

Turning to the point estimates, we throw out the first 50 iterations as burn-in and obtain the following values:

$$\hat{\beta} = \begin{bmatrix} -.89 & .02 & -.04 & .85 \\ -1.67 & .10 & -.32 & .71 \\ -1.49 & -.15 & -.12 & 3.15 \\ .63 & .03 & -.03 & 1.80 \\ .84 & -.29 & .19 & 1.67 \end{bmatrix}, \tag{13}$$

$$\hat{R} = \begin{bmatrix} 1 & .30 & .07 & .22 & .04 \\ .30 & 1 & -.06 & -.07 & -.01 \\ .07 & -.06 & 1 & .52 & .16 \\ .22 & -.07 & .52 & 1 & .43 \\ .04 & -.01 & .16 & .43 & 1 \end{bmatrix}, \tag{14}$$

$$\hat{\gamma}_{4,2} = 2.56, \tag{15}$$

$$\hat{\gamma}_{5,2} = 3.20. \tag{16}$$

Considering that we are estimating 37 parameters in five dimensions from a sample size of only 429, the posterior means are relatively close to the true value.

6.3 Breast Cancer Prevention Study

In this section, we analyze some of the results of the pilot study discussed in Section 2. In particular, we are interested in modeling the outcome of six survey questions based on individual demographic (see Appendix A for the survey information). Data from 289 completed questionnaires were used in the analysis. Although the main purpose of the pilot study was to examine features of the survey, there is also interest in studying the preliminary results of the regarding attitudes and knowledge about breast cancer and the treatment tamoxifen. We use the variable selection procedure to evaluate the significance of the demographic information on the respondent's choices.

Figure 7 shows the Markov chains for the data analysis. The model was run for 1000 iterations. As before, the chains converge quite quickly from the starting values. In this case, this fact is particularly satisfying as there is a relatively small amount of data when we consider the number of choices for each response and the amount of parameters that are estimated.

To answer any questions about demographics, first consider Figure 8. Clearly, there are two models that occur significantly more often than any others. The most frequent model, at about 16%, features no significant variables. Slightly less significant, at about 12%, is a model featuring one significant variable: a family history of breast cancer has a significant effect on the whether the respondent's perception of the likelihood of breast cancer compared to the average woman. The

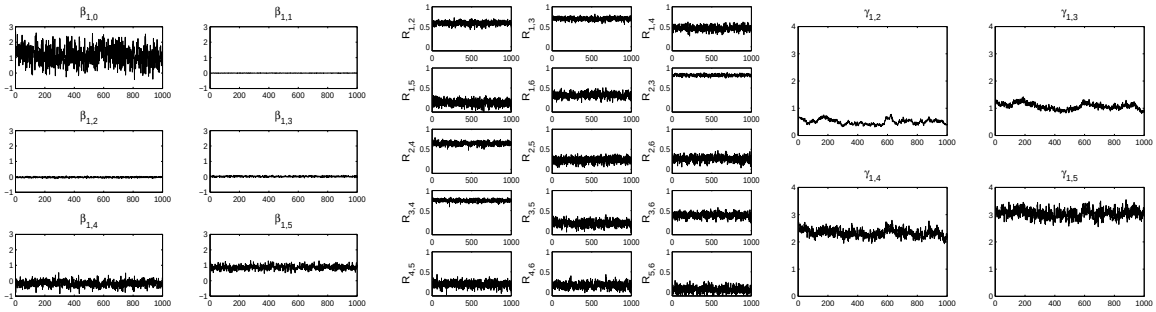

Figure 7. Markov chains of β_1, the correlations, and γ_1 for the breast cancer survey data.

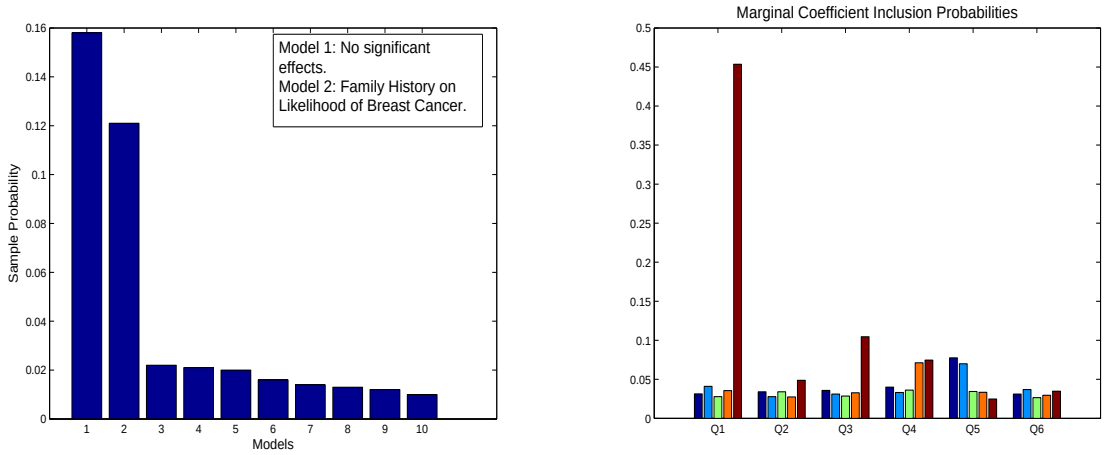

Figure 8. Right: Sample probabilities for the ten most frequent models from the breast cancer survey data. Models 1 and 2 are significantly more frequent than the others. Left: Marginal inclusion probabilities for each covariate.

posterior mean for this coefficient is about .87, indicating that a respondent with a family history of breast cancer thinks she is more likely than the average woman. This is certainly a reasonable result to expect.

Alternatively, we can look at the marginal coefficient inclusion probabilities, also show in figure 8. The relative frequencies are important here as the prior probability of inclusion can play a large role in the posterior probability. Six coefficients stand out in this plot. Family history of breast cancer is potentially significant in questions 1, 3, and 4, all questions related to the perceived likelihood of developing breast cancer. Knowing someone else with breast cancer is also potentially significant in question 4. Age and education level are potentially significant in a respondent's

assessment of the possible impact of breast cancer. The posterior estimate of β is given by:

$$\hat{\beta} = \begin{pmatrix} 1.08 & 0 & -0.03 & 0.02 & -0.18 & 0.87 \\ 0.91 & 0 & -0.02 & 0 & 0.13 & 0.38 \\ 0.78 & -0.01 & 0.01 & -0.01 & 0.19 & 0.53 \\ 1.33 & -0.01 & -0.02 & 0.03 & -0.36 & 0.47 \\ 2.90 & -0.01 & 0.05 & 0 & -0.06 & -0.05 \\ 0.55 & 0 & -0.03 & 0 & -0.04 & 0.18 \end{pmatrix}.$$

Finally, consider the posterior estimates of the correlation matrix:

$$\hat{R} = \begin{pmatrix} 1.00 & 0.59 & 0.70 & 0.47 & 0.13 & 0.32 \\ 0.59 & 1.00 & 0.82 & 0.65 & 0.23 & 0.27 \\ 0.70 & 0.82 & 1.00 & 0.75 & 0.17 & 0.39 \\ 0.47 & 0.65 & 0.75 & 1.00 & 0.19 & 0.17 \\ 0.13 & 0.23 & 0.17 & 0.19 & 1.00 & 0.05 \\ 0.32 & 0.27 & 0.39 & 0.17 & 0.05 & 1.00 \end{pmatrix}.$$

There are a few large positive correlations among the responses. The questions on perceived likelihood are highly correlated with each other, as perhaps is expected. The question about worry is also correlated with this group suggesting that those who perceive a high likelihood of developing breast cancer worry about it more than others (or worrying leads to a high perception of likelihood of development).

7 Conclusion

In this article, we developed a parameter expanded MCMC procedure for the MVP model. The estimation procedure uses draws of an unstructured covariance matrix from an inverse Wishart distribution and re-scales to get a draw from the desired covariance matrix. The formulation is helpful because it overcomes the identifiability constraints of the MVP model and also because it allows for easy MCMC implementation. Furthermore, the parameter expansion formulation facilitates fast convergence of the Markov chain.

8 Acknowledgements:

The authors are grateful to Dr. Angela Fagerlin of the Program for Improving Health Care Decisions at the University of Michigan for providing us with the data on Application 1. Nair's research was supported in part by NSF Grant DMS0204247 and National Cancer Institute Grant P50 CA 101451.

A Breast Cancer Survey Selection

A.1 Survey Questions

1. Compared to the average woman (your age and in your health), what are your chances of developing breast cancer in your lifetime?

   0 (Much less than average) - 1 - 2 - 3 (Same as the average) -4 - 5 - 6 - 7 (Much higher than average)

2. How worried are you that you will develop breast cancer in your lifetime?

   0 (Not worried at all) - 1 - 2 - 3 - 4 - 5 (Extremely worried)

3. If you were to choose not to take tamoxifen, how likely do you think you would be to get breast cancer in your lifetime?

   0 (Not at all likely) - 1 - 2 - 3 - 4 - 5 (Extremely likely)

4. If you were to choose to take tamoxifen, how likely do you think you would be to get breast cancer in your lifetime?

   0 (Not at all likely) - 1 - 2 - 3 - 4 - 5 (Extremely likely)

5. If you were to develop breast cancer, how much of an impact do you think it would have on the quality of the rest of your life?

   0 (Not very much impact at all) - 1 - 2 - 3 - 4 - 5 (A lot of impact)

6. Do you think that taking tamoxifen to prevent breast cancer is worth tamoxifen's potential health problems for you?

   0 (No, for me it is definitely not worth the potential health problems) - 1 - 2 - 3 - 4 - 5 (Yes, for me it is definitely worth the potential health problems)

A.2 Demographic Information

1. Age

2. What is the highest level of education you have completed?

   1. None; 2. Elementary School; 3. High School; 4. Trade School; 5. Some college but no degree; 6. Associate degree; 7. Bachelors degree; 8. Masters degree; 9. Doctoral or Professional Degree

3. What is your annual household income before taxes?

   1. Less than $10,000; 2. $10,001 - $25,000; 3. $25,001 - $40,000; 4. $40,001 - $60,000; 5. $60,001 - $80,000; 6. $80,001 - $100,000; 7. More than $100,000

4. Have you ever known anyone who has been diagnosed with breast cancer?

5. Has anyone in your immediate family ever been diagnosed with breast cancer?

REFERENCES

Agresti, A. (1984), *Analysis of Ordinal Categorical Data*: John Wiley & Sons.

Ashford, J. R. and Sowden, R. R. (1970), "Multi-variate Probit Analysis," *Biometrics*, 26(3), 535–546.

Bock, R. D. and Gibbons, R. D. (1996), "High-Dimensional Multivariate Probit Analysis," *Biometrics*, 52(4), 1183–1194.

Chib, S. and Greenberg, E. (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85(2), 347–361.

Chipman, H. A. (1998), "Fast Model Search for Designed Experiments with Complex Aliasing," in *Quality Improvement Through Statistical Methods*, ed. B. Abraham, Birkhauser.

Edwards, Y. D. and Allenby, G. M. (2003), "Multivariate Analysis of Multiple Response Data," *Journal of Market Research*, XL, 321–334.

Fay, J. W. J. (1957), "The National Coal Board's pneumoconiosis research," *Nature*, 180, 309.

Friedman, D., Hansen, M. H., Nair, V., and James, D. (1997), "Model-free estimation of some yield metrics in integrated circuit manufacturing," *IEEE Trans on Semiconductor Manufacturing*, 10(3).

Gassmann, H. I., Deák, I., and Szánti, T. (2002), "Computing Multivariate Normal Probabilities: A New Look," *Journal of Computational and Graphical Statistics*, 11(4), 920–949.

George, E. I. and McCulloch, R. E. (1993), "Variable Selection Via Gibbs Sampling," *Journal of the American Statistical Association*, 88(423), 881–889.

Gupta, A. K. and Nagar, D. D. (2000), *Matrix Variate Distributions*: Chapman & Hall/CRC.

Hansen, M. H., Nair, V., and Friedman, D. (1997), "Monitoring Wafer Map Data from Integrated Circuit Fabrication Processes for Spatially Clustered Defects," *Technometrics*, 39(3).

Johnson, V. E. and Albert, J. H. (1999), *Ordinal Data Modeling*: Springer.

Kutsyy, V. (2001), *Modeling and inference for spatial processes with ordinal data*, unpublished Ph.D. dissertation, University of Michigan.

Likert, R. (1932), "A Technique for the Measurement of Attitudes," *Archives of Psychology*.

Liu, C. (2001), "Discussion on The Art of Data Augmentation," *Journal of Computational and Graphical Statistics*.

Liu, C., Rubin, D., and Wu, Y. (1998), "Parameter Expansion to Accelerate EM: The PX-EM Algorithm," *Biometrika*, 85(4), 755–770.

Liu, J. S. and Wu, Y. (1999), "Parameter Expansion Scheme for Data Augmentation," *Journal of the American Statistical Association*, 94, 1264–1274.

McCullagh, P. (1980), "Regression Models for Ordinal Data," *Journal of the Royal Statistical Society, Series B*, 42(2), 109–142.

McCullagh, P. and Nelder, J. A. (1989), *Generalized Linear Models*: Chapman & Hall/CRC.

McCulloch, R. E. and Rossi, P. E. (1994), "An Exact Likelihood Analysis of the Multinomial Probit Model," *Journal of Econometrics*, 64, 217–228.

Ochi, Y. and Prentice, R. L. (1984), "Likelihood Inference in a Correlated Probit Regression Model," *Biometrika*, 71, 531–543.

Ware, J. H., Dockery, D. W., Spiro III, A., Speizer, F. E., and Ferris Jr., B. G. (1984), "Passive Smoking, Gas Cooking, and Respiratory Health in Children Living in Six Cities," *American Review of Respiratory Disease*, 129, 366–374.